DOCUMENT RESUME

ED 186 045 IR 008 810

AUTHOR Carifio, James; Biron, Ronald
TITLE Collecting Sensitive Data Anonymously: Further Findings on the CDRGP Technique.
SPONS AGENCY National Inst. on Alcohol Abuse and Alcoholism (DHEW/PHS), Rockville, Md.
PUB DATE Apr 80
GRANT H84-AA-01242-04
NOTE 48p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 7-14, 1980)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Confidentiality; *Data Collection; *Methods; Questionnaires; Tables (Data)

ABSTRACT

This report of research findings describes a confidentiality-preserving data collection method called the Context-Determined, Rule-Generated, Pseudonym (CDRGP) technique, a specifically constructed set of questions based upon personal information known to the respondent (but to few others) which each respondent answers whenever data are gathered; the respondent's answers to the questions generate a probabilistically unique string of letters, or code identifier, that allows the data being collected to be tied to previously collected data. The report describes CDRGP in full, explains the research methodology involved in three separate experiments, and details the findings. References are provided. (FM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED186045

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Collecting Sensitive Data Anonymously: Further Findings on the CDRGP Technique

James Carifio, Massachusetts Board of Regional Community Colleges

Ronald Biron, Marathon House

The need for a simple, confidentiality-preserving, data-collection technique capable of generating correlational data across multiple time-points has been a long standing methodological problem. The CDRGP data collection technique was devised to solve this problem. Previous experiments showed that the technique worked well with seventh to twelfth grade students and that responses were connectable across multiple time-points in 96% of the instances. The present experiments showed that students would respond to alcohol consumption questions using the technique and that no significant response biases were introduced through use of the technique.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

James Carifio

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The research presented in this paper was supported in part by a grant from NIAAA, grant no. 2, H84 AA 01242-04

Paper presented at the annual meetings of the American Education Research Association in Boston, Massachusetts, April 7 to April 14, 1980.

IR008430

Sensitive data on topics such as alcohol and drug use are typically collected by self-report, paper-and-pencil techniques. Since the data being collected is sensitive, researchers usually try to protect the privacy of the respondents doing the reporting by disguising their identities. Theoretically, enhanced response validity is supposed to be gained by disguising respondent identities (Hyman, 1970). Whether or not enhanced response validity is actually obtained, however, is somewhat of an irrelevant matter, since some form of identity protection is now required by law, even for non-sensitive data (see Michael and Weinberger, 1977 for details)[1].

The techniques presently available for use in disguising respondent identities are (1) the pre-coded questionnaire (King, 1971), (2) the anonymously responded to questionnaire (Guilford, 1957), (3) the randomized response technique (Warner, 1965), and (4) the contamination technique (Boruch, 1972). Although each of these techniques has its own advocates, all have severe limitations associated with them in terms of (a) their actual mechanics of use and (b) the data analyses that can be performed if they are actually employed.

---

[1] Whether identity protection enhances the validity of the data obtained more than other kinds of responding conditions, it should be noted, is strictly an empirical question. The argument that anonymity reduces response bias (lying, exaggeration, concealment, and/or evasiveness) has yet to be convincingly demonstrated. There is a reasonable amount of evidence to suggest that response bias is a psychological characteristic that manifests itself to some degree in the testing situation regardless of the conditions of responding (Clawar, 1975). Identity protection does quite obviously enhance the validity of the response obtained from many individuals in many situations but more work needs to be done with some of the newer technology available such as Bell's (1975) voice analysis polygraph device before the hypothesis can be accepted as a generalizable fact.

Briefly, the major limitations of the above techniques are as follows (see Carifio and Biron, 1978 for fuller details):

(1) Responses made to data collecting stimuli using the above techniques cannot be connected over time, and, in some instances, within the same time set due to the use of randomization procedures. Consequently, correlational or effects analysis as they have come to be known in statistical literature cannot be done with data collected by these techniques unless one accepts some form of cohort analysis procedure which is an extremely risky and tenuous business (Ulric, 1978);

(2) One of the response techniques (the randomized response technique) is limited in that only 3 or 4 questions at most can be administered at one time; and,

(3) Three of the techniques are extremely cumbersome to employ and to administer, particularly with relatively young subjects (namely, techniques 1, 3, and 4).

Although the pre-coded questionnaire is only subject to limitation (3) above, it should be noted that pre-coded questionnaires and master files can be subpoenaed by authorities and that there are documented instances of this occurring with drug related data (Flynn, 1976).[1]

In an effort to deal with the above mentioned difficulties, the present authors developed a relatively simple data collection technique called the Context - Determined, Rule - Generated, Pseudonym (CDRGP) Technique.

---

[1] The pre-coded questionnaire also has other limitations in addition to those cited above. The necessary fore-knowledge aspect of the pre-coded questionnaire makes it almost impossible to employ in a great number of longitudinal and school-based research situations. In schools, students are entering and leaving the system at different grade levels on a yearly basis. In many situations, whether a person is an experimental subject, or what treatment is received, is completely dependent upon the person's pattern of voluntarism over some time period of interest. Ignoring the anonymity issue, pre-coding is simply not feasible in the field in a great many instances.

This data collection technique is a specifically constructed set of questions that are based upon personal information known to the respondent but few others, which each respondent answers whenever data are collected (see Table 1). The respondent's answers to the specifically constructed set of questions (called the standardized responding rule) generates a probabilistically unique string of letters (called the code identifier) that allows the new data being collected to be tied to previously collected data, thus making all responses for each subject who is present on both data collecting occasions "connectable" over time.

TABLE 1

STANDARDIZED RESPONDING RULE NOW USED
IN ALL PAST AND PRESENT STUDIES OF THE
CDRGP TECHNIQUE BY THE PRESENT AUTHORS

CASPER CODE NAME

Please read each statement below carefully and write the correct letter for each question on the left-hand line. PLEASE PRINT CLEARLY.

_____First letter of your middle name (If no middle name, write "z")

_____First letter of the month you were born in

_____First letter of your sex: Male or Female

_____First letter of the name of your street

_____First letter of your mother's first name (If unknown, write "x")

Questions to be answered placed here with directions

In addition to the feature that subject mortality rates from data collecting occasion to data collecting occasion may be probabilistically assessed with the CDRGP technique, a feature not present in the other techniques mentioned at the beginning of this paper, the following points about the CDRGP technique should be noted by the reader:

(1) The CDRGP technique is context-determined because the standardized responding rule can be adapted or varied to fit the age level of the subject, the sensitivity of the data being collected, or various other important data collecting parameters of the data collecting situation at the researcher's discretion. The CDRGP technique, therefore, is not a particular operational solution, but rather a generic approach to the problem; namely, the responding rule that is actually used in a particular situation is constructed by the researcher to meet the requirements of that particular situation.

(2) As previously indicated, the "content" of the "sentences" chosen to be included in the standardized responding rule should be private, personal information that is easily remembered and highly stable over time; namely, items should be "questions" like "number of older brothers" rather than "number of brothers" which can be subject to rapid change over time in a school age population.

(3) The "syntax" of the "sentences" chosen to be included in the standardized responding rule should be cognitive operations that are easy to follow and perform by the subjects doing the responding. This point is particularly true of younger subjects and in terms of having a responding rule that is employable by the entire responding population (namely, students with special needs).

(4) Any coding system, as it is well known, is always ultimately breakable given enough resources, associated information, and computer time. Additionally, a given coding schema may have hidden flaws, or hidden flaws in a particular situation. Extreme care, therefore, should be taken relative to safeguarding associated contextual information in the experimental situation and in terms of evaluating a given standardized responding rule against a given situation for glaring flaws.

(5) The matching (connecting) of "unique" letter strings (code identifiers) across time points is, quite obviously, based upon probabilistic assumptions about (a) the random production of duplicate letter strings in a given population (or sample) on a given occasion, and (b) standardized responding element errors across time points in a given responding population or sample. A great deal will be said about these two particular points later in this paper, as these code identifier matching probabilities have been misunderstood by some users of the technique who have been throwing away usable data, but it should suffice for now to point out that:

(a) the theoretical probability of random duplicates for the simple responding rule given in Table 1 is roughly 1 in 300,000 (on the average);

(b) the maximum theoretical error rate relative to connecting up 4 of 5 letter matches for the simple responding rule given in Table 1 is roughly 1 in 11,000 (taking into account letter position and neglecting the use of associated information or exact letter probability distributions);

(c) the exact probabilities of particular 5 letter random duplicates (and 4 of 5 letter duplicates) are dependent upon the exact letter distributions of each rule element included in the rule used and the population sampled; and,

(d) making false code identifier connections in a given situation will, at most, introduce an extremely low additional error factor into the error factor used for analysis, given both the maximum and minimum matching probabilities that typically tend to be operational in a given situation.

In a word, objective reality is objective reality, and false matches at worst will produce outlier data points that will increase the error terms for analysis at a faster rate than the cell means. Additionally, it should be pointed out that the number of subjects one ends up debating about is, in our experience, usually less than 5% of the sample in question, and at false error matching rates of typically 20,000 to 1. If one finds such beta error rates intellectually or theoretically unacceptable, one can, of course, always alter one's standardized responding rule to produce more acceptable error rates. Probabilistic matching, however, is the name of the game with the technique we have devised.

(6) Conceptually, the CDRGP technique (to use a computer analogy) can be thought of as being somewhat akin to defining each person at "object time" through the generation of appropriate "code" prior to the "data run". This aspect of the technique, and the essentially cognitive view of the respondent it presents, grounds the technique psychologically in a cognitive learning theory rather than in more behaviorally oriented learning or measurement paradigms; and,

(7) Further theoretical grounding for the CDRGP technique may be found in what is called public key cryptology theory (Hellman, 1979). Briefly, public key cryptology systems are based on principles of (a) computational intractability relative to having enough computation power available to unravel what has been coded, (b) trap door one-way coding functions that do not allow the original "text" to be completely and unambiguously recovered from the "code text" and (c) two coding keys; namely, an en-coding key and a decoding key. The relationship of the CDRGP technique to public key cryptology theory will be made more explicit at the end of this paper, but the connections are relatively obvious and straightforward as soon as one realizes that the positioning of the elements in the standardized responding rule could be varied in the overall data

string from occasion to occasion with dummy elements inserted, and then one would need two keys to break the lock on the data itself prior to analysis time.[1]

Lastly, it should be noted that students who have used the CDRGP technique have tended to call the technique, the "code name" technique. Additionally, many researchers have told us that the label "CDRGP", which was meant to be meaningful via technique modeling, initially communicated very little to them about the technique itself. Consequently, in deference to the opinions of others, we will employ the term "code name" rather than"CDRGP" technique wherever possible in order to facilitate better communication about the technique itself relative to potential future users.

Initial Findings

We have been generating experimental evidence about the CDRGP technique over the past three years using the simple procedure given in Table 1, as this simple responding rule has proven to be sufficient in terms of collecting data about alcohol attitudes and knowledge from students in grades 3 to 12. Our primary research focus has been on younger students.

We initially field-tested the code name technique using alcohol knowledge and attitude items with 259 seventh to twelfth grade students in a larger inner suburban community in eastern Massachusetts.

The methodology we used in these experiments was to purposefully choose one "shy", "bright", "slow", and "problem" student from each exposed classroom after the technique was implemented by the classroom teacher in the course of evaluating an

---

[1] One could, or course, also encrypt the data responses themselves which would add a third and fourth key to the process (if one were really worried about records being subponaed, or wanted to participate in the ultimate mystification of the research process). However, rather than going to such lengths as these, the present writers would suggest seeking First Amendment legal protection as a more viable problem solution strategy.

alcohol education curriculum on a pre, post, and interim basis. We then randomly chose a male or female student from the same classroom to "balance" each student purposefully chosen. These students were then ramdomly interviewed by an adult male or 18 year old male interviewer according to a standardized interview protocol. No interviewer difference in results was found.

In general, the code name technique was found in these experiments to be easy to perform, and to have a high degree of face validity and credibility with the students who participated, the vast majority of whom (87%) reported understanding why the technique was being employed; namely, to protect their identities and to make program evaluation possible (see Carifio and Biron, 1978 for details). Students also reported that they believed that the letter strings would be difficult to unravel and attach to a name, and that they liked the fact that the coding rule generated a "meaningless string" of letters; namely, that the production of pronouncable letter strings was rare.

The multiple time-point data were connectable in 96% percent of the cases (over a two to three week interval), and only two students reported that they would not respond to a personal alcohol consumption questionnaire using the code name technique[1]. The outstanding questions from these experiments, however, were (1) would students actually respond to alcohol consumption questions using the code name technique, (2) would the consumption data collected map back into other data that was over a year old, and (3) would the use of the technique bias the

---

[1]This short-term connectability finding was further replicated at roughly the same percentage levels in two subsequent studies involving 287 students and 626 students over a one to four week period (see Carifio, Biron, and White (1978) for further details.)

data or results obtained as compared to anonymously collected consumption data. Answering these three questions and a few others, therefore, were the primary goals of the experiments to be reported on here.

Methodology and Results

Since the results of three different experiments are to be reported here, methodology and results will be presented together purely due to organizational convenience. It should also be noted that the same responding rule as given in Table 1 was used in each of the experiments reported here, and that supplemental data (i.e., grade and teacher) were used in some instances to cross-validate results and make definite code name matches.

The alcohol consumption (drinking behavior) questionnaire used in all of the studies reported here was a modified version of the consumption questionnaire developed by the Research Triangle Institute (1975) for the National Institute of Alcohol Abuse and Alcoholism (NIAAA). This instrument (originally consisting of 119 consumption, knowledge, attitude and associated information items) will be described more fully in experiment 3, as (1) it was reduced by approximately 40% in terms of items relative to the instrument used in experiment 1 (which was approximately a 50% reduction of the original instrument), and (2) experiment 3 is the experiment in which the direct comparison of data collection methods was made, thereby making the second reduced version of this instrument the pertinent instrument to describe.

Experiment 1

The first experiment conducted was designed to test the code name technique under actual field performance conditions. A sample of 1,132 grade 7 to grade 12 students was drawn from an inner suburban school system in eastern Massachusetts

and another sample of 1,100 grade 7 to grade 12 students was drawn from a second community in close proximity to the first to respond to the first scaled down version of the Triangle Research Institute's alcohol consumption questionnaire. Random classrooms were used as the sampling unit in both communities, as school officials in both communities felt that students might feel stigmatized if they were randomly drawn from classrooms and assembled in one place for testing. Approximately 24% of the students available at each grade level in both communities were sampled by the classroom sampling procedure employed, and all students tested in the samples drawn were tested during the month of May in 1978.

All students in both communities responded to the alcohol consumption questionnaire employed using the code name technique. The omnibus questionnaire used, it should be noted, was the criterion measure of an experiment comparing two different alcohol education approaches, and the samples drawn were mixed in terms of students who had prior exposure to the code name technique and those who did not. In terms of prior exposure, the data that was available on exposed students was 12 to 17 months old and was confined to a sample of students who had been tracked for evaluation purposes on a pre and post test basis in the curriculum evaluation study that was conducted in the previous year.

The alcohol consumption questionnaire was administered to all students in their intact classrooms by two adult male administrators who explained to all students the purpose of the data collection activities and why the code name procedure was being employed. Lastly, it should be noted t at this experiment was not conducted under the direction of the senior author of this paper, but rather under the direction of another principal investigator who replaced the senior author on this large scale evaluation project when the senior author moved on to other professional activities. Experimentally, the consequences of this particular fact are:

(1) Direct access by the present senior author to all of the data that were collected and were available (for the purpose of carrying out several analyses that might have been interesting) was prohibited; and,

(2) The principal investigator in charge of the project at the time the consumption data were collected was under a two month reporting deadline in terms of a wide variety of important analytical work that had to be carried out with a limited budget.

Consequently, a number of "gliches" in the analyses and results to be reported here are attributable to these two factors.

According to the junior author of this paper, who was one of the two test administrators in the first experiment, no student in either community refused to take the alcohol consumption questionnaire using the code name technique (N=2,232). Approximately 11 students in the sample initially had reservations about responding to the alcohol consumption questionnaire using the code name technique, but these reservations were dispelled with further explanation of the technique. Although these findings are somewhat straightforward, interpretation relative to these findings will be postponed until after reporting the results of experiment 3, as the results of both experiments are needed to make a complete summary statement.

Experiment 2

The second experimental question that was posed at the beginning of this paper was, "to what degree would the consumption data collected in experiment 1 map back into data from the original curriculum evaluation study that was from 12 to 17 months old?" Answering this question required making assumptions about (a) what happens to various student populations over a 12 to 17 month period and (b) on any particular testing day in order to (c) generate a priori expected availabilities of students to be matched in the 1978 samples.

In order to simplify the presentation of data relevant to answering the question posed, the reader should note at the outset that the matching search

that was conducted was limited by the principal investigator in charge of the project at this point to those students in community 1 (the experimental community) who had been tracked on a pre and post test basis in the curriculum evaluation study that had been conducted 12 to 17 months earlier. This target group, therefore, was comprised of eighth, ninth, and twelfth graders only in terms of the 1978 omnibus consumption measure sample, as these grades were the only grades in community 1 that had received exposure in the prior year (see Table 2 for details).

TABLE 2

BASIC COMPUTATIONAL DATA FOR GENERATING EXPECTED CODE NAME MATCHES IN 1977-78 CONSUMPTION SURVEY

GRADE	1976-78 TOTAL N	1977-78 TOTAL N	TRACKED N	ATTRITION RATE	AVERAGE ATTANDANCE RATE	PERCENTAGE SAMPLED
8	942	904	206	4.0%	87.5%	24.0%
9	997	948	180	4.9%	87.5%	24.0%
11	933	877	144	6.0%	86.9%	24.0%
TOTALS	2872	2729	530	5.0%	87.2%	24.0%

The next step in the matching expectancies generation process was developing some reasonable assumptions about what happens to students over a 12 to 17 month basis in an inner suburban school system and on any particular testing day. School officials in the community in question were able to proyide us with attri-

TABLE 3

COMPUTATION FORMULAE FOR GENERATING
EXPECTED CODE NAME MATCHES

MODEL A: No attrition, no absenteeism etc.:

$$PS \times TN = EM$$

MODEL B: Attrition factored in:

$$PS \times (TN - AxTN) = EM$$

MODEL C: Attendence factored in:

$$PS \times ADAR \times TN = EM$$

MODEL D: Attrition and Attendance factored in:

$$PS \times ADAR \times (TN - A \times TN) = EM$$

Where:

PS = Percentage Sampled

TN = Tracked N (Students)

ADAR = Average Daily Attendance Rate

A = Attrition Rate

E = Expected Matches

---

tion rate data[1] and average daily attendance rate data[2] for each of the grades in question.

Since the percentage of students that were sampled at each grade level was known (see Table 2), one now only had to use the computational formulae given in Table 3 to compute out expected matching frequencies under 4 different sets of assumptions. These assumptions were:

(a) expected matching frequencies assuming no attrition or absenteeism (Model A) which represents the ideal carry forward and maximum presence case;

(b) expected matching frequencies factoring in attrition only (Model B) which assumes that the losses due to daily attendance are trivial;

(c) expected matching frequencies factoring in attendance only (Model C) which assumes that losses due to attrition occurred for non-tracked students only; and,

(d) expected matching frequencies factoring in both attendance and attrition (Model D) which assumes that both factors are important and affect all students equally.

All of the above models and the computational formulae used, it should be noted, assume in one way or another equal attrition and absenteeism rates and strict random sampling which was not the case in terms of the data in question. These particular assumptions add an unknown error factor to the expected matching frequency values generated which simply could not be evaluated. This unknown

---

[1] Attrition rate here does not mean (necessarily) a student dropping out of school for good. It means a student leaving the school system in question to attend another school system within the community or in another community. In the community in question, there were 3 separate school systems into which a student could transfer; namely, parochial, vocational, or home-based systems.

[2] Although actual attendance rates on the day testing actually occurred should theoretically have been used, this was not possible due to (a) the testing being spread over the month of May and (b) classroom data being non-recoverable. How actuals would be computationally entered into equations is a somewhat moot point, but daily attendance varied from 10% to 15% (excluding explainable outliers) and this fact should be noted.

error factor, however, cannot be very large due to the nature of the methodology employed and the matching models developed.

Once the expected frequency values were generated, appropriate computer files were run against one another to search out code name matches. The matching process at each grade level, it should be noted, was divided into two "passes". The first pass consisted of 5 for 5 letter matches and the second pass consisted of 4 of 5 letter matches, taking into account letter position. More will be said about this second pass later when matching probabilities are discussed in more detail. Lastly, it should be noted that confining the search process to a particular grade level (namely, working grade level by grade level) increases all false matching probability ratios by a multiplicative factor of 5, and is what is meant by the phrase "narrowing the field of search through use of associated information".[1]

Table 4 summarizes the results of the observed to expected code name matching experiment. As can be seen from Table 4, observed to expected matching percentages ranged from a low of 75.6% to a high of 96.4%, depending upon the matching model one is considering.[2]

The present authors consider Model C to be the most conservative yet most realistic model for assessing expected to observed matches. As can be seen from

---

[1] It should be noted that each computer file is pre-assessed for duplicates within the file, within the field of search so that it is possible to either (a) eliminate subjects a priori or (b) use additional information to distinguish duplicate code names, see pages 17-20 for further details.

[2] From page 10.

Table 4

PERCENTAGES OF EXPECTED TO OBSERVED
CODE NAME MATCHES BY MATCHING ASSUMPTIONS

	GRADES			
EXPECTED CODE NAME MATCHES	8	9	11	TOTAL
1. Model A: No attrition, absenteeism etc.	49	43	34	126
2. Model B: Attrition factor in	47	41	32	121
3. Model C: Attendence factor in.	43	38	30	108
4. Model D: Both attrition and attendance factored in.	42	36	28	106
OBSERVED CODE NAME MATCHES				
1. First Pass: 5 for 5 letter matches.	38	34	26	98
2. Second Pass: Pick up on 4 of 5 letter matches	40	35	27	102
CODE NAME MATCHING PERCENTAGES				
A. First Pass:				
1. Model A	77.5%	79.1%	76.5%	77.8%
2. Model B	80.8%	82.9%	83.1%	81.0%
3. Model C	88.0%	89.5%	86.7%	90.7%
4. Model D	90.5%	94.4%	92.8%	92.4%
B. Second Pass:				
1. Model A	81.6%	81.4%	79.4%	80.9%
2. Model B	85.1%	85.4%	84.4%	84.3%
3. Model C	93.0%	92.1%	90.0%	94.4%
4. Model D	95.2%	97.2%	96.4%	96.2%

Table 4, the expected to observed match rate for Model C is 94.4%. This matching rate, pragmatically, must be considered to be nothing less than excellent, particularly considering that (1) the time lag between data points is 12 to 17 months, (2) the standardized responding rule used was not specifically designed with tracking over this length of time in mind, (3) the most unstable rule element in the standardized responding rule used tended to be street code, and (4) students in the system in question could have "disappeared" in a year's time into the parochial, vocational, or alternative school systems that were also operating in this community. Experimentally, the present writers believe that the longitudinal code name matching question originally posed has been strongly answered in the positive. The technique does in fact produce, and is quite capable of producing, qualitative longitudinal data matching as the results presented in Table 4 clearly indicate.[1]

Code Name Matching Probabilities

Since the question of "code name matching probabilities" continually comes up whenever researchers are exposed to the code name technique and matching process for the first time, this issue will be addressed here rather than elsewhere, as it naturally follows from the data and results just presented. Prior to addressing this commonly raised issue, however, it must be pointed out that the issue itself is, with very little doubt, a non-issue for the following reasons:

---

[1] It is the present writers' belief that further analysis of the full data base that was available for analysis in this project would simply further corroborate the findings presented here and that such further analysis would constitute little more than a testing of the theoretical limits of the particular standardized responding rule used. For example, the analysts did not test missing code names against prior grades to see if missing students might have stayed back. It is, therefore, our belief that the question posed has been strongly answered in the affirmative.

(1) When one examines Tables 5 and 6, which give the "average" probabilities of random code name duplicates and random 4 of 5 letter matches under 5 different letter distribution assumptions for the standardized responding rule given in Table 1, one will quite clearly see that the probability values under discussion are so small fractionally that one is talking about, at most, an extremely small number of subjects in the typical real world experimental situation; namely, 1 in 100,000 to 1 in 7,000 subjects by the most conservative of estimating procedures. The overwhelming vast majority of code names generated by the simple standardized responding rule given in Table 1 are unique, as may be readily seen from the data presented in Table 5.

(2) Since the issue at hand concerns the probabilities of random duplicate code names and random 4 of 5 letter matches, what one is actually talking about, in fact, is introducing random error into the data by making false matches. Random error primarily affects the error terms used for significance testing, so, at worst, in terms of relatively rare "events" one is talking about biasing significance tests in the conservative direction.

(3) Given the probability values that will be presented in Tables 5, 6, and 8, it should, with minimum reflection, become relatively clear to the reader that the research benefits to be gained through use of the code name technique far outweigh the error risks in question, particularly relative to the analytical and methodological alternatives currently available; and,

(4) The issue of "code name matching probabilities" is, in fact, really a non-issue, since one may pre-analyze the initial sample of data collected for duplicate code names that cannot be distinguished by any other associated information and eliminate these subjects from the data file, thus eliminating the problem.[1]

With the above points in mind, therefore, we will now focus more specifically on the question of "code name matching probabilities".

As previously mentioned, the exact probabilities of random duplicates, false matches, and n-x of n position letter matches is contextually a function of the standard responding rule used and the exact letter distributions of the rule

---

[1] One obviously, moreover, does not have to accept 4 of 5 letter matches, and one can, of course, always raise one's alpha levels to gain comfort. The number of subjects that one will end up discarding, however, with the ultimate in methodological purity will, from our experience, be extremely small.

elements in the rule used in a given population. Therefore, in order to supply discussion at this point, the present writers will focus on the standardized responding rule given in Table 1 and some reasonable assumptions about the nature of letter distributions for this rule in different populations.

Table 5 presents the probabilities of random duplicate code names occurring under five different assumptions about code letter distributions in a given population. Case 1 in Table 5 represents the "completely theoretical" case; namely, case 1 gives duplicate code name probabilities under the assumption of equally probable letter distributions across all rule elements which we know not to be the case due to naming preferences, street name truncations in given communities, and several other factors similar to the factors just mentioned.

Case 2 in Table 5 is an actually observed case in terms of exact letter probability distributions for one class of seventh graders on whom data was collected in 1977 (see Tables 7 and 8), as these were the only data available to the senior author for developing an exact observed probability distribution for illustration purposes (see Carifio and Biron, 1978 for the actual raw data). Cases 3 and 4 are logical degradations of case 2 (which is highly degraded relative to case 1 as it is based on data from a single, relatively homogeneous classroom), and case 5 is akin to collecting data in a single classroom in a small, ethnically homogeneous community where family size tends to be large (namely, the brother-sister problem).[1]

---

[1] If one were faced with collecting data in a case 5 situation, one would obviously not use the standardized responding rule used for experimental purposes in this study but another responding rule that would more adequately deal with the duplicate code name and 4 of 5 letter match problems. Alternatively, one could, of course, simply add elements to the rule used in this study to solve the problem.

TABLE 5

PROBABILITIES OF RANDOM DUPLICATE CODE NAMES UNDER DIFFERENT ASSUMPTIONS ABOUT CODE DISTRIBUTIONS IN A POPULATION

Response Rule Element	Case 1	Case 2	Case 3	Case 4	Case 5
1	1/26	1/14	1/10	1/10	1/6
2	1/8	1/8	1/8	1/6	1/6
3.	1/2	1/2	1/2	1/2	1/2
4.	1/26	1/15	1/10	1/10	1/6
5.	1/26	1/13	1/10	1/10	1/6
DUPLICATE CODENAME PROBABILITES	$\frac{1}{281,216}$	$\frac{1}{43,680}$	$\frac{1}{12,000}$	$\frac{1}{8,000}$	$\frac{1}{1,596}$
ADJUSTED DUPLICATE PROBABILITY[1]	$\frac{1}{562,432}$	$\frac{1}{87,360}$	$\frac{1}{24,000}$	$\frac{1}{16,000}$	$\frac{1}{2,592}$

[1] Adding one other binary bit of information.

TABLE 6

PROBABILITIES THAT FOUR LETTER CODE NAME MATCHES ARE NOT TRUE MATCHES UNDER DIFFERENT ASSUMPTIONS ABOUT CODE DISTRIBUTIONS IN A POPULATION

	Nature of Error	Case 1	Case 2	Case 3	Case 4	Case 5
1.	Error in 1st letter; remainder match	$\frac{1}{10,816}$	$\frac{1}{3,120}$	$\frac{1}{1,200}$	$\frac{1}{800}$	$\frac{1}{216}$
2.	Error in 2nd letter; remainder match	$\frac{1}{35,152}$	$\frac{1}{5,460}$	$\frac{1}{2,000}$	$\frac{1}{2,000}$	$\frac{1}{432}$
3.	Error in 3rd letter; remainder match	$\frac{1}{140,608}$	$\frac{1}{21,840}$	$\frac{1}{6,000}$	$\frac{1}{4,000}$	$\frac{1}{648}$
4.	Error in 4th letter; remainder match	$\frac{1}{10,816}$	$\frac{1}{2,912}$	$\frac{1}{1,200}$	$\frac{1}{800}$	$\frac{1}{216}$
5.	Error in 5th letter; remainder match	$\frac{1}{10,816}$	$\frac{1}{3,360}$	$\frac{1}{1,200}$	$\frac{1}{800}$	$\frac{1}{216}$
6.	maximum error rate	$\frac{1}{10,816}$	$\frac{1}{3,360}$	$\frac{1}{1,200}$	$\frac{1}{800}$	$\frac{1}{216}$
7.	adjusted[1] maximum error rate	$\frac{1}{21,632}$	$\frac{1}{6,720}$	$\frac{1}{2,400}$	$\frac{1}{1,600}$	$\frac{1}{432}$

[1]Addition of one binary bit of information.

TABLE 7

EXACT LETTER DISTRIBUTIONS BY CODE NAME RULE
FOR ONE CLASS OF SEVENTH GRADE STUDENTS (N-33)

	Code Name Rule						Code Name Rule				
Letter	1	2	3	4	5	Letter	1	2	3	4	5
A	4	4	0	1	1	N	0	3	0	3	1
B	1	0	0	3	2	O	0	3	0	0	0
C	1	0	0	5	4	P	0	0	0	2	2
D	2	4	0	0	2	Q	0	0	0	0	0
E	2	0	0	1	1	R	1	0	0	1	1
F	1	4	18	0	0	S	0	4	0	1	3
G	1	0	0	2	0	T	0	0	0	0	0
H	1	0	0	4	0	U	0	0	0	0	0
I	0	0	0	2	0	V	1	0	0	0	0
J	3	8	0	2	4	W	0	0	0	1	0
K	0	0	0	0	0	X	0	0	0	0	3
L	1	0	0	2	1	Y	0	0	0	0	0
M	6	3	15	3	8	Z	8	0	0	0	0
				TOTAL N.:			33	33	33	33	33
			TOTAL LETTERS USED:				14	8	2	15	13

TABLE 8

EXACT PROBABILITIES OF FALSE CODE NAME MATCHES FOR ONE CLASS OF SEVENTH GRADE STUDENTS (N-33)

	Most Common Letters	Least Common Letters
Random duplicate	$\frac{1}{850}$	$\frac{1}{2,609,902}$
Error in 1st letter; remainder match	$\frac{1}{206}$	$\frac{1}{26,353}$
Error in 2nd letter; remainder match	$\frac{1}{206}$	$\frac{1}{79,062}$
Error in 3rd letter; remainder match	$\frac{1}{463}$	$\frac{1}{395,307}$
Error in 4th letter; remainder match	$\frac{1}{129}$	$\frac{1}{26,353}$
Error in 5th letter; remainder match	$\frac{1}{206}$	$\frac{1}{26,353}$
Maximum error	$\frac{1}{129}$	$\frac{1}{26,353}$

To fully understand the values presented in Table 5, the reader should look at Tables 7 and 8 first. What can be seen from Tables 7 and 8 is that different code names in a real-world population have different probabilities of duplication; namely, code names comprised of commonly occurring letters have a higher probability of duplication than code names comprised of less common letters. The probability of duplicate code names occurring, therefore, is a distributed variable which is dependent in turn on the sample or samples one is dealing with at any given time. What should be reasonably clear from the latter two points is that (1) the probability values given in Tables 5 and 6 are in fact more akin to "average" probabilities, and (2) one could sophisticate one's matching decisions (or matching process) through use of rarity values if one felt that the matching situation warrented or demanded it. With these two points in mind, Tables 5 and 6 will be discussed.

As can be seen from Table 5, the "average" probability of random code name duplicates ranges from roughly 1 in 300,000 to 1 in 1,600, depending upon the case number in question. As previously mentioned, one is talking about very few subjects indeed when one talks about random duplicate code names.

Table 6 presents the probabilities of random 4 of 5 code letter matches under the 5 different assumptions about code letter distirbutions given in Table 5. As can be seen from Table 6, the maximum false matching rate error that would occur from connecting 4 of 5 letter matches for the standardized responding rule used in Table 1 ranged from 1 in 11,000 to 1 in 216 (on the average).[1] Again, using case 2 as the best available estimating case, the maximum

---

[1] The probability of 4 of 5 letter matches is not 1 in 3 as some collegues have claimed, but rather is as the odds given in Table 6 indicate. The 1 in 3 claim is derived, most probably, from computing 4 of 5 letter matches in any order using Pascal's triangle, or standard combinatorial probability formulae. If letter position is utilized in mathcing, however, the false matching error rate odds are not 1 in 3 but are the odds given in Table 6.

false-matching error rate that would be likely to occur would be roughly 1 in 3,000, neglecting any consideration of code name rarity or use of associated information.[1] It is our belief that this error rate constitutes an extremely low level of risk, particularly when one considers that one is talking about random error and its overall effects on data analysis.

It is also our belief that the data that has been presented in Tables 5 to 8 clearly shows that the code name matching probability problem is truly a non-problem, and that the research benefits to be gained through use of the code name technique far outweigh the error risks involved, particularly when compared to the analytical and methodological alternatives that are presently available.

Experiment 3

As previously stated, experiment 3 was designed to answer the question, "Would use of the code name technique bias the consumption data collected or the results obtained as compared to anonymously collected consumption data?" Bias here, it should be noted, is defined comparatively in terms of one data collection technique as opposed to another, and not in terms of the accuracy of the data obtained by either technique (see page 1 of this paper for further details on this point).

In experiment 3, the manner in which the above question was addressed was as follows:

---

[1] It should be noted here that we are talking about 4 of 5 letter matches between two data points and not three data points as given in experiment 1. With additional data points, all false-matching odds decrease expotentially.

Table 9
SUMMARY OF ITEM RESPONSE TOPOLOGIES USED IN
EXPERIMENT THREE BY VARIABLE NAME AND IDENTIFICATION NUMBER

Variable Number	Variable Abbreviation	Variable Name	NRC[1]	Response Values
1.	NDCTS	Number of drinking occasions per annum	8	0, 1, 4, 12, 36, 52, 156, 365
2.	NHIGH	Number of times high per annum	8	0, 1, 4, 12, 36, 52, 156, 365
3.	NDRUNK	Number of times drunk per annum	8	0, 1, 4, 12, 36, 52, 156, 365
4.	AVEMO	Number of drinking occasions in average month		0, 1, 2, 3, 4, 12, 16, 30
5.	AVEND	Average number of drink per drinking occasion	12	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 15+
6.	GND	Greatest number of drinks on one occasion	10	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 15+
7.	NSDH	Total Number of drinking situations in which subjects got drunk or high	9	0, 1, 2, 3, 4, 5, 6, 7, 8,
8.	FDRINK	Father's drinking	4	1, 2, 3, 4,
9.	MDRINK	Mother's drinking	4	1, 2, 3, 4,
10.	LDRINK	Wish parent(s) drank less	3	1, 2, 3,
11.	FACTS	Alcohol fact items (10)	3	0, 1, 2, : Percent of total wrong
12.	ATTS	Alcohol attitude items (10)	3	0, 1, 2, : Percent of total negative

[1]Number of Response Categories

(1) In a third community adjacent to community 1 in experiment 1, 120 seventh to ninth grade students were randomly selected and randomly assigned to one of two treatment conditions; namely, filling out an alcohol consumption questionnaire using the code name technique or filling out the same questionnaire anonymously. Seventh to ninth grade students were used as the treatment allocation unit in this experiment as non-willingness to respond tends to occur in these grades, and a great deal is known about how the cell means of these grades should behave statistically. The sample drawn was balanced in terms of sex, and all students were naive subjects relative to the use of the code name technique.

(2) Two random back-up name lists were developed to deal with the absenteeism problem. Five students in the code name responding condition and 7 students in the anonymous condition, who were originally drawn to participate in experiment 3, were absent from school on the day of testing. All of these students were replaced with appropriate students of the random back-up lists.

(3) The alcohol consumption measure used in experiment 3 was a scaled-down version of the Triangle Research Institute (TRI instrument developed for NIAAA (TRI, 1973). This omnibus consumption measure is described in detail in Table 9. As previously mentioned, this omnibus consumption measure was a scaled-down version of the scaled-down version of the TRI instrument used in experiment 1. As can be seen from Table 9, this second scaled-down version of the TRI instrument was comprised of 15 personal consumption items, 3 items on parental drinking, 10 alcohol fact items, and 10 alcohol attitude items, plus grade and sex indicators where needed.[1] In most instances, verbatim items were used from the TRI insturment to gain comparability with known norms, and the items included were chosen to relect the types of alcohol consumption questions that typically tend to be asked.[2]

---

[1] Alcohol fact and attitude items were scored in terms of percentages of wrong responses as this has been found to be the best type of scoring procedure for these items (see Carifio and Biron, 1978).

[2] As is well known, there are no good alcohol consumption measures, due in part to the unit-per-unit-time measurement problem involved and due to the fact that measurement specialists have not spent a great deal of time working on the problem. To the best of the senior author's knowledge, there are no test-re-test relationship estimates on the type of consumption items employed in this experiment, but all error would be constant due to the nature of the question and the experimental situation.

(4) As can be noted from Table 9, each variable in the consumption measure employed could be analyzed in either of two ways; namely, through use of a Responding Condition by Grade, by Sex (2 x 3 x 2) analysis of variance design treating each variable as an interval variable, or non-parametrically through use of Chi Squares and Kendall Tau U coefficients relative to treating each variable and the values of each variable as nominal or ordinal data elements. The data collected was analyzed both parametrically and non-parametrically, but it should be noted that Glass et al (1972) and Marks (1974) have shown that F-ratios are extraordinarily robust in terms of violation of ANOVA assumptions and that the analysis of variance design presents the clearest picutre of the data.

Lastly, it should be noted that experiment 3 was conducted under the direction of the senior author of this paper.

Each treatment condition in experiment 3 included 2 students who would not participate; namely, 2 students who refused to fill out the consumption questionnaire. In the code name technique conditions, one seventh grade female and one eighth grade male refused to fill out the consumption questionnaire. In the anonymous responding condition, one seventh grade male and one eighth grade female refused to fill out the consumption questionnaire. No student who refused to respond was replaced in the data collection sample for reasons that may not at first be apparent to the reader, but which will be explained below.

It is a reasonably well known fact that some students will refuse to fill out sensitive data questionnaires under any responding conditions. The results obtained above, therefore, are not atypical relative to the experiences of other researchers in this area. At another level, there would also seem to be a great deal of validity to the views of school officials cited in experiment 1; namely, that some students might feel stigmatized if they were randomly drawn from classrooms and assembled in one place for testing. At a completely different level, however, there is also an additional factor operating that one might tend to overlook.

TABLE 10

SUMMARY OF UNIVARIATE F-RATIOS FOR SURVEY ITEMS USED IN EXPERIMENT THREE

		VARIABLES											
SOURCE	df	1	2	3	4	5	6	7	8	9	10	11	12
RC	1	0.0	0.0	0.1	0.2	0.0	0.0	0.2	0.0	0.0	0.0	0.3	0.0
G	2	3.6[a]	5.6[b]	2.7[a]	3.8[a]	10.8[b]	5.8[b]	12.1[b]	0.1	0.3	12.4[b]	36.9[b]	10.0[b]
S	1	5.5[a]	5.3[a]	5.4[a]	5.4[a]	6.4[b]	7.5[b]	8.6[b]	0.6	0.1	8.6[b]	21.9[b]	20.2[b]
RC x G	2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.7	0.1	0.1	0.0
RC x S	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0
G x S	2	0.2	0.4	0.5	0.2	0.1	0.0	0.3	0.0	0.1	0.3	0.7	0.3
RC x G x S	2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.7	0.7
Error	108												
TOTAL	115												

Significant F-Ratios Underlined:: [a]Significant at .05 [b]Significant at .01

KEY: RC = Responding Condition: CDRGP or Anonymous Response

G = Grade (7, 8 and 9)

S = Sex (male and female)

TABLE 11

MEAN NUMBER OF DRINKING SESSIONS PER ANNUM
BY RESPONDING CONDITION, GRADE AND SEX

	CDRGP TECHNIQUE				ANONYMOUS TECHNIQUE			
	Males		Females		Males		Females	
GRADE	n	$\bar{x}$	n	$\bar{x}$	n	$\bar{x}$	n	$\bar{x}$
7	10	38.8	9	8.5	9	41.0	9	7.3
8	9	77.8	10	24.0	10	76.9	10	24.9
9	10	94.9	10	65.4	10	93.7	10	64.6
Total	39	70.8	39	32.6	39	70.5	39	32.3

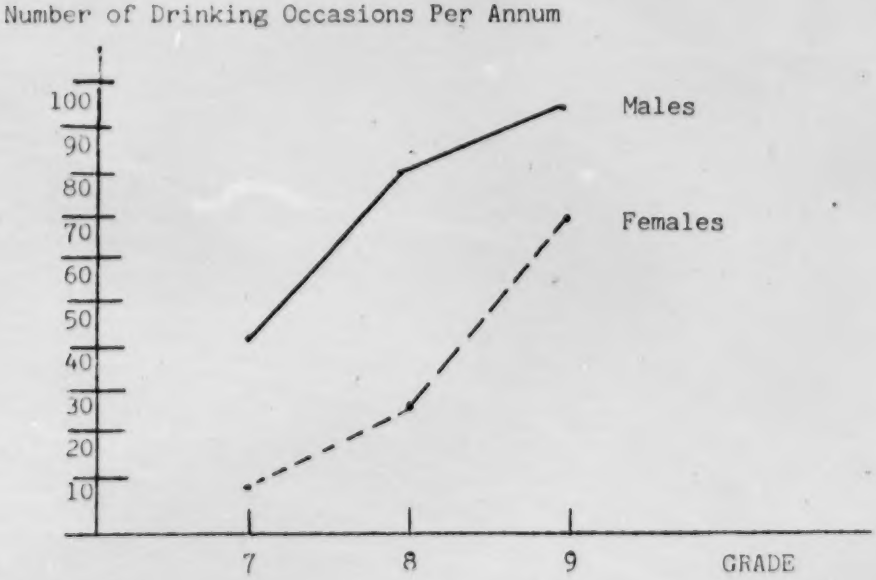

Figure 1: Number of Drinking Occasions Per Annum
By Grade and Sex

In each administrating condition, the senior author of this paper informed students of their right to not participate in the data collecting activities that were about to occur; namely, of their right to remain silent, as it were, which is not exactly an everyday experience in the lives of seventh to ninth grade students. The degree to which the behavior of the 4 students in question represented "testing out" behavior we cannot say. However, we can say that the right exercised should be respected, and be actively shown to be respected, regardless of what motivating conditions might be prevailing. More will be said on this point and about these results later in this paper.

In terms of the central analytical question in experiment 3, Table 10 presents a summary of the univariate F-Ratio observed for each of the 12 response variables on which data was collected according to the Responding Condition by Grade by Sex design exmployed.

As can be seen from Table 10, no significant differences between responding conditions were observed on any of the 12 variables scrutinized. Significant grade and sex effects were observed where they should have been observed (i.e., variables 1 to 7 and 10 to 12) and were not observed where they should not have been observed (i.e., variables 8 and 9).[1]

In terms of observed cell means, Table 11 and Figure 1 present the results for variable 1, Number of Reported Drinking Sessions per Annum. As can be seen from Table 11 and Figure 1, the cell means and trends on this variable are very close to TRI norms for the northeast, with the exception of the sharper catch-up rate of females which has also been observed elsewhere (see White and Biron, 1978 for details). The same type of results as present in Table 11 and Figure 1

---

[1] A multi-variate analysis of these data would obviously reveal the same results as given in Table 10.

were also observed on variables 2 to 7 and 10 to 12.

Table 12 presents a summary of Chi Square analyses for the dependent variables and main factors given in Table 10, and Table 13 presents a summary of Kendall Tau U coefficients. In addition to providing a non-parametric re-analysis of the data collected in experiment 3, Table 12 (the Chi Square analyses) also tests the degree to which students employed the potential range of responding categories available on each variable in the same manner, and Table 13 (the Kendall Tau U analyses) also tests this same category spread factor taking the ordinality of the data into account.

This category spread factor is important because researchers who analyze consumption data often use the categorical extremes to analyze response. For example, the responses of abstainers are typically compared to those of problem drinkers. Tables 14 and 15, therefore, present the number of abstainers and problem drinkers that were identified in experiment 3 by Responding Condition, Grade, and Sex.

As can be seen from Tables 14 and 15, no significant differences between responding conditions were observed in terms of number of abstainers and problem drikers identified. Grade differences and total sex differences were observed as they should have been, which simply corroborates the finding presented in prior tables. All analyses of differences in Tables 14 and 15 were carried out, it should be noted, through use of Fischer Z-Tests for proportions.

In summary, no significant differences between responding conditions were observed in experiment 3 in terms of the 12 analytical variables scrutinized.

Table 12

SUMMARY OF CHI-SQUARE RATIOS FOR SELECTED SURVEY ITEMS USED IN EXPERIMENT THREE

Variable	Responding Condition	Grade	Sex
1. NDOPA	0.3 (df=7)	15.7 (df=14)	24.3 (df=7)[b]
2. NHIGH	0.5 (df=7)	18.9 (df=14)	10.7 (df=7)
3. NDRUNK	0.1 (df=7)	7.2 (df=14)	11.5 (df=7)
4. AVEMO	0.3 (df=7)	12.4 (df=14)	23.5 (df=7)[b]
5. AVEND	5.2 (df=11)	38.0 (df=22)[b]	7.6 (df=9)
6. GND	12.7 (df=11)	37.5 (df=22)[b]	18.1 (df=11)
7. NSDH	1.6 (df=8)	25.2 (df=18)[a]	13.0 (df=8)
8. FDRINK	3.8 (df=3)	1.2 (df=6)	0.9 (df=3)
9. MDRINK	1.4 (df=3)	0.6 (df=6)	0.2 (df=3)
10. LDRINK	0.1 (df=2)	0.2 (df=4)	0.2 (df=2)

a. Significant at .05 level

b. Significant at .01 level

These findings replicated King's (1971) findings of no significant differences between pre-coded and anonymously responded to questionnaires on drug usage. The comparison of the code name technique to the anonymous technique on drug usage items would obviously be both an interesting and worthwhile experiment, but the present writers are fairly confident at this point that the outcome

Table 13

SUMMARY OF KENDALL'S TAU U COEFFICIENTS FOR SELECTED SURVEY ITEMS USED IN EXPERIMENT THREE

Variable	Responding Condition	Grade	Sex
1. NDOPA	-0.01	0.24**	-0.25**
2. NHIGH	-0.00	0.26**	-0.27**
3. NDRUNK	0.01	0.16*	-0.24**
4. AVEMD	-0.02	0.23**	-0.28**
5. AVEND	-0.03	0.35**	-0.25**
6. GND	0.00	0.28**	-0.29**
7. NSDH	-0.01	0.38**	-0.27**
8. FDRINK	0.04	-0.03	0.09
9. MDRINK	0.02	0.03	-0.02
10. LDRINK	0.01	0.04	0.01

*Significant at the 05 level

**Significant at the 01 level

of such an experiment will be similar to experiment 3 reported here (see Carifio and Biron, 1978 for further details).[1] Lastly, although no significant differences

[1]Babst, Uppal, and Schmerdler (1978) have found that substance abuse tends to approximate a Guttman scale; namely, those subjects who tend to be abusing one substance tend to be involved with others, and the more substances used, the more likely it is that the usage is frequent, recent, and concurrent. In addition to our 1978 finding of all except 2 students telling us they would provide drug use information using the code name technique, Babst, Uppal, and Schmerdler's findings tell us that there were more than likely drug users in the sample we tested.

TABLE 14

NUMBER OF ABSTAINERS BY RESPONDING CONDITION, GRADE AND SEX

	CDRGP	TECHNIQUE	ANONYMOUS	TECHNIQUE
	Males	Females	Males	Females
GRADE	N	N	N	N
7	3	4	2	5
8	1	4	2	5
9	1	2	2	2
Total	5	10	6	10

TABLE 15

NUMBER OF PROBLEM DRINKERS BY RESPONDING CONDITION, GRADE AND SEX

	CDRGP	TECHNIQUE	ANONYMOUS	TECHNIQUE
	Males	Females	Males	Females
GRADE	N	N	N	N
7	2	0	2	0
8	3	2	3	1
9	3	2	4	2
Total	8	4	9	3

in refusal to respond rates were observed between the two responding conditions tested in experiment 3, something was potentially learned about the general conditions under which sensitive data might have to be collected from younger students due to a number of mitigating factors; namely, stigmatization effects and novelty effects relative to human subjects legislation. Both of these factors are worthy of further qualitative and quantitative investigations.

Further Findings

Over the past two years, the present writers have used the code name technique to gather sensitive data in a wide variety of situations other than the types of situations reported on up to this point in this paper. We have used the technique to gather information from court-mandated youth in special treatment programs for driving-related alcohol offenses, from children of alcohol troubled families in special treatment groups, from teenaged peer leaders while in training and after training, and from students in intensive alcohol awareness study groups. In all of these situations, we experienced exceedingly few problems with the CDRGP technique, and we uncovered a number of additional benefits other than those that have already been cited here and elsewhere.[1]

For example, treatment group directors continually told us about their feelings of relief from knowledge in terms of gathering data from clients even after the treatment intimacy factor took hold; namely, it was a relief not to have to know "traceable particulars" about each client, but a boon to be able to know

---

[1]Court mandated youth, children from alcoholic families, and peer leaders on occasion refused to provide information using the code name technique. In most instances, however, these subjects refused to provide the information requested in any shape or form. The most common reason cited for refusal was a fear that treatment directors might be able to figure out who the respondents were due to the smallness of the group. The delicate irony of this particular fear will become apparent in a moment.

results and trends in general terms.[1] This same feeling of protection was also experienced in these situations by the present writers; namely, although we could stitch the data together and analyze it over time, we simply did not know what data came from whom in particular, which was an extremely comforting feeling (to say the least), especially when it came to small things like having a friendly cup of coffee with treatment participants. Probabilistic models of knowing, therefore, have a number of additional benefits other than just respondent identity protection, and it is our experience in these latter situations that have given us a great deal of confidence in the code name technique as it relates to a number of small group reactivity issues and the more humane side of the research and evaluation process.

Summary and Conclusions

Approximately four years ago, we undertook several research and program evaluation tasks that made us directly confront the problem of collecting sensitive data in a way that protected each respondent's identity. After a thorough review of the literature, we found that all of the identity protection methodologies available were severely limited in that subject responses to stimulus items could not be connected over time and in some instances within the same data collecting occasions. After considerable reflection and experimentation, we developed the initial prototype of a generalized methodology that was capable of overcoming the "anonymous-but-yet-connectable-over-time-problem". We named

---

[1] Although use of the code name technique sometimes becomes a little artificial in relatively small groups after the treatment intimacy factor takes hold, we have always maintained use of the technique so that neither participants nor data clerks know what information is coming from whom.

the generalized methodology developed the Context-Dependent, Rule-Generated, Pseudonym Technique (CDRGP Technique) in order to communicate via modeling the nature of the technique and its products. With time and the accumulation of data, the technique became re-named the "code name" technique by those who actually employed it.

We initially tested the code name technique very cautiously and conservatively in a series of highly controlled pilot experiments that employed peer and adult interviews (see Carifio and Biron, 1978). The code name technique in these initial experiments was found to be easy to do and to have a high degree of face validity and credibility with the students (259 seventh to twelfth graders) who participated in these experiments, the vast majority of whom (87%) reported understanding why the technique was employed; namely to protect their identity and make program evaluation possible. These students also reported that they believed that the letter strings would be difficult to unravel and attach to a name, and that they liked the fact that the coding rule generated a "meaningless string" of letters; namely, that the production of a pronouncable letter string was rare.

The multiple time-point data collected were connectable in 96% of the cases over a two to three week interval[1], and only two students reported that they would not respond to a personal alcohol consumption questionnaire using the code name

---

[1]This short-term connectability finding was further replicated at roughly the same percentage levels in two subsequent studies involving 287 students and 626 students over a one to four week period (see Carifio, Biron, and White (1978) for further details).

technique. These initial experiments, moreover, also generated a great deal of soft, qualitative information about the technique, its fundamental nature, and sensitivities and insensitivities to a variety of factors that should be reviewed by any reader who is interested in using or further researching the technique. These same experiments, however, were also limited in that they purposefully did not attempt to collect actual alcohol consumption data (thereby keeping the experimental response ante low), and they did not address or answer a number of other important questions about the code name technique.

In the various experiments reported in this paper it was found that students would readily employ the code name technique when providing truly sensitive data about both personal and parental alcohol consumption. It was also found, however, that subtle aspects of the testing situation can produce a refusal to respond that did not differ in terms of the two responding methodologies employed; namely, the code name or the anonymous responding technique. This finding most definitely warrents further investigation in terms of a wide variety of both testing techniques and testing situations.

It was additionally found in experiment 3 that use of the code name technique introduced no observable statistical bias in the alcohol consumption data collected as compared to anonymously provided alcohol consumption information, and that the cell means and trends of the data obtained under both responding conditions examined were not significantly different from cell means and trends in a number of other studies (which means that this finding is highly generalizable population-wise).

The question of whether new data collected using the code name technique could be mapped back into longitudinal data that was 12 to 17 months old was addressed empirically in the present series of studies reported here and answered positively. Using Expentancy Model C (factoring in daily attendance only) the

expected to observed matching rate was found to be 94.4% overall in terms of the 8th, 9th, and 11th grade students who were in the "tracked group" in 1977 and the survey sample in 1978.

The question of duplicate code names and 4 of 5 letter matching probabilities was addressed in detail in this paper and judged to be essentially a non-issue for a variety of reasons, the most outstanding of which were (1) the problem could be eliminated as a problem on an a priori basis if necessary (as well as by other techniques), and (2) the probability values of case 2 (of the 5 demonstration cases developed) were so low as to constitute arguments about an extremely few subjects in any given sample, ignoring the effects of the introduction of random error and risk-benefits aspects of the issue in question. A qualifier will be added to this "non-issue" conclusion when our thoughts on "directions for further research" are presented; however, at the practical level of conducting qualitative field-based research, we firmly assert this "non-issue" position as a fundamental and undeniable conclusion. The "average" probability values in the highly degraded cases presented in this paper are simply too high to assert otherwise.

In a great number of other experiments with the code name technique (many of which pre-dated experiments 1 and 2 reported in this paper), we found that the code name technique could be used with very few problems to gather a wide variety of sensitive data from court-mandated youth, children of alcoholic families, teenaged peer leaders, and students in intensive alcohol awareness groups. In these various experiments with small groups, we also uncovered an additional benefit of the code name technique which we called the "relief from knowledge" factor. This factor was judged to be an important characteristic of the code name technique relative to both reactivity issues and the day to day conduct of

real-world program activities.[1]

Lastly, at the beginning of this paper, we related the code name technique to the work that is being done in public key cryptography theory (Hellman, 1979), since the conceptual connections are relatively straightforward and easy to make. Although we believe that the essentially cognitive nature and original framework of the code name technique can never be overlooked, public key cryptography can be used to link various aspects of the code name technique to a much broader conceptual and mathematical area of inquiry, which might prove to be quite beneficial in terms of constructing more complex and subtler standardized responding rules than the simple prototype standardized responding rule we have employed in our initial studies.

In looking at the body of knowledge we have built up in the past four years regarding the code name technique, we have come to some conclusions about where further research and development work on the technique should go from here. These lines of further inquiry are as follows:

---

[1] We also found that the code name technique is not free of unintended outcomes; namely, many students wanted to do all of their school work using the technique which could most definitely prove to be a non-trivial problem.

(1) In the past four years, we have come to realize that the code name technique is a relatively sophisticated type of research technology due primarily to its generic character. We have tried in our various writings to provide potential users with as much detailed knowledge of the technique's subtlies as possible, particularly in terms of unforeseen traps that one can back into and how to back out of them. One area where we have not provided a great deal of detailed guideline is in the area of code name matching, although this paper addressed this issue in more detail than anything else we have written and has outlined the process sufficiently enough at the general level so that it can be followed by anyone familiar with file-search programming (see Lewis and Papadimitrou, 1978, and Graham, 1978, for further details). It is this very nature of the code name technique, however, that is in need of further development so that the technique can be used by relatively unsophisticated users; namely, one line of further research that needs to be conducted is on heuristic matching algorithms and computer programs that will produce maximal matching outcomes relative to the standardized responding rule used to generate the input. One can consume a great deal of computer time with large files and a bump-and-grind program, and work needs to be done on this heuristic algorithm problem if only to gain methodological standardization.

(2) One impetus to the development of the CDRGP technique was some trial research that we did which attempted to match up anonymously responded to alcohol consumption questionnaires through biographical profile items. Such questionnaire matching is theoretically quite possible but was thwarted due to (a) profile item construction and (b) lack of a good computerized algorithm to do the job. Such lines of pursuit could, quite legitimately, be called "unobtrusive matching techniques", and are extraordinarily interesting theoretically when one begins to think about personality inventories, attitude inventories, and strings of achievement test items on a pre and post-test basis; namely, if one had a well-formed theory of the nature of the intervention delivered, what post-tests probabilistically should be matched up with pre-tests (relative to "decoding" the nature of the intervention delivered), given initial pre-test levels.

(3) After considerable reflection, we are not completely sure about how long-range, cost-effective follow-up would be done on subjects using the code name technique once these subjects left the general environment of the research or evaluation situation. Further reflective and experimental research most definitely needs to be done on this point as long term follow-up is increasingly being emphasized; and,

(4) Generalized investigations into different standardized responding rule types for various kinds of information and populations need to be conducted in terms of generating some empirical guidelines for the "matching probabilities problem" relative to "small town", "case study" type situations where standardized responding rule construction could be problematic but most definitely surmountable

Although the above mentioned lines of further research would be extremely helpful in advancing knowledge about the code name technique and a number of other issues associated with the technique, enough is presently known about the technique and its empirical behavior to draw some very fundamental conclusions.

The need for a simple, confidentiality-preserving, data collection technique capable of generating correlation data across multiple time points has been a long standing methodological problem. The CDRGP technique is indeed simple and confidentiality-preserving and was shown in the present set of experiments to work under actual field conditions with no apparent statistical bias being introduced as compared to anonymously collected alcohol consumption information.[1] Methodologically sound research, therefore, is now without doubt practically possible in the sensitive data area through use of the code name technique. The practical utility of the code name technique is simply undeniable.

Lastly, the present authors have experienced certain reservations about the CDRGP technique from time to time in terms of both its simplicity and slightly artificial if not prosaic character. On other occasions, almost in an impish plea for proper perspective, we have retreated into wry whimsy as one must maintain above all a sense of humor about this whole field of research relative to sensitivities, data, and methodologies. All of these feelings, however, quickly disappear as soon as we begin to reflect again on the nature of the issue being

---

[1] Students have reported that they would be willing to answer drug related questions using the CDRGP Technique (see Carifio and Biron, 1978), a finding we doubt that experimentation will deny.

addressed; namely, informational privacy and the rights of human subjects are not trivial issues and neither is the issue of adequate effects analysis.

References

Babst, J., Uppal, B., and Schmerdler, K. Relationship of Youths' attitude to substance use in New York State. Journal of Alcohol and Drug Education, 23, 2; 1978, 24-37.

Boruch, R. Relations among statistical methods for assuring confidentiality of social research data. Social Science Research 1, 1972, 403-414.

Carifio, J. and Biron, R. Collecting sensitive data anonymously: the CDRGP technique. Journal of Alcohol and Drug Educaiton 1, Winter, 1978, 26-52.

Carifio, J., Biron, R. and Sullivan, D. Selection findings on the CASPAR alcohol education program. National Institute on Alcohol and Alcohol Abuse, 1978, ERIC ED 177 169 TM008 934.

Flynn, P. Personal Communication, May 26, 1976.

Glass, G., Peckham, P., and Sanders, J. Consequences of failure to meet assumptions underlying the analysis of variance and co-variance. Review of Educational Research, 3, 42, Summer, 1972, 237-288.

Graham, D.L. The combinatorial mathematics of scheduling. Scientific American, 238, 3, March, 1978, 124-133.

Guilford, J. Psychometric Methods, McGraw-Hill, New York, 1954.

Hellman, M.E. The mathematics of public key cryptography. Scientific American 241, 2, August, 1979, 138-145.

King, F. Anonymous versus identifiable questionnaires in drug usage surveys. The American Psychologist, 25, 1070, 982-985.

Lewis, H.R., and Papadimitrou, C.H. The efficiency of algorithms. Scientific American, 238, 1, January, 1978, 96-109.

Marks, E. Analysis of variance technique for unbalanced designs. Review of Educational Research, 3, 44, Summer, 1974, 351-364.

Michael, J., and Weinberger, J. Federal restrictions on educational research: protection of research participants. Educational Researcher, 6, 1, January, 1977, 1-8.

Triangle Research Institute. Adolescent drinking, behavior, attitudes, and Correlates. Triangle Research Institute, Research Triangle Park, No. 27709, TRI, P0237 - 891, 1973.

Ulric, R. Cohort Analysis, Sage Press, 1978.

Warner, S. Randomized response: a survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 60, 1965, 63-69.

White, R., and Biron, R. Final Report on the CASPAR Alcohol Education Program. National Institute of Alcohol Abuse and Alcoholism, Rockville, Maryland, 1978.